Genetic Resources

ORIGINAL ARTICLE

Genetic Resources (2023), 4 (8), 55–63
DOI: 10.46265/genresj.EMOR6509
https://www.genresj.org
ISSN: 2708-3764

A public mid-density genotyping platform for alfalfa (*Medicago sativa* L.)

Dongyan Zhao[a], **Katherine Mejia-Guerra**[a,b], **Marcelo Mollinari**[c], **Deborah A Samac**[d], **Brian M Irish**[e], **Kasia Heller-Uszynska**[f], **Craig T Beil**[a] **and Moira J Sheehan**[*,a]

[a] *Breeding Insight, Cornell University, Ithaca, 14853, NY, USA*
[b] *Sarepta Therapeutics, Cambridge, 02142, MA, USA*
[c] *Campus Box 7609, North Carolina University, NC, Raleigh, 27695, USA*
[d] *Plant Science Research Unit, USDA-ARS, St. Paul, 55108, MN, USA*
[e] *Plant Germplasm Introduction and Testing Research Unit, USDA-ARS, Prosser, 99350, WA, USA*
[f] *Diversity Arrays Technology, ACT 2617, Bruce, Australia*

Abstract: Small public breeding programmes have many barriers to adopting technology, particularly creating and using genetic marker panels for genomic-based decisions in selection. Here we report the creation of a DArTag panel of 3,000 loci distributed across the alfalfa (*Medicago sativa* L.) genome for use in molecular breeding and genomic insight. The creation of this marker panel brings cost-effective and rapid genotyping capabilities to alfalfa breeding programmes. The open access provided by this platform will allow genetic data sets generated on the marker panel to be compared and joined across projects, institutions and countries. This genotyping resource has the power to make routine genotyping a reality for any breeder of alfalfa.

Keywords: Alfalfa, amplicon-sequencing, plant breeding, DArTag genotyping, microhaplotype

Citation: Zhao, D., Mejia-Guerra, K., Mollinari, M., Samac, D. A., Irish, B. M., Heller-Uszynska, K., Beil, C. T., Sheehan, M. J. (2023). A public mid-density genotyping platform for alfalfa (*Medicago sativa* L.). *Genetic Resources* 4 (8), 55–63. doi: 10.46265/genresj.EMOR6509.

Introduction

Molecular breeding techniques have been used for nearly four decades to enhance and speed breeding efforts for major staple food crops like tomato, maize and barley (Tanksley, 1983; Helentjaris *et al*, 1985; Feuerstein *et al*, 1990; Hasan *et al*, 2021). Over time, these techniques have been augmented with high-quality phenotypic data to perform genome-wide association studies (GWAS) and genomic selection and prediction, further fueling breeding for quantitative or complex traits (Eathington *et al*, 2007; Lorenzana and Bernardo, 2009; Heffner *et al*, 2009). While these achievements are significant, many crop species grown for human consumption and livestock feed are still unable to apply these techniques in breeding efforts. Many breeders would like to adopt molecular breeding tools and techniques, but sometimes doing so is hampered by large barriers-to-entry challenges. The range of barriers and how surmountable they are, varies from species to species and is impacted by species-specific challenges in logistics, technical know-how, biology and the growing environment.

Alfalfa (*Medicago sativa* L.) is the most widely grown perennial forage crop worldwide (Undersander, 2021). In the United States, it was the fourth most cultivated crop in 2021 with an estimated direct value of US$11.6 billion (Putnam and Meccage, 2022) and ranked first among forage crops planting area with a total of 14.9 million acres in 2022 (https://www.nass.usda.gov/). Alfalfa is a key nutritional component for dairy and beef production because it contains a high amount of crude protein, provides dietary fibre needed to

[*]Corresponding author: Moira J Sheehan (moirasheehan@cornell.edu)

Received: 20.07.2023 | Accepted: 26.09.2023 | Published online: 13.10.2023

maintain rumen health, and is an excellent source of vitamins and minerals. In addition, it is unparalleled as a component of sustainable agricultural systems because of its ability to fix nitrogen, protect water quality, interrupt pest and pathogen cycles in annual crops, and improve soil carbon storage (Fernandez *et al*, 2019). Alfalfa is adapted to different growth environments and depending on location and management, is highly persistent.

As a highly heterozygous outcrossing autotetraploid species, alfalfa features a predominant pattern of random chromosome pairing during meiosis. The four sets of chromosomes add a layer of complexity to genotyping endeavours. Traditional SNP marker systems, primarily designed for diploid species, often fall short when applied to alfalfa due to their inability to identify allelic dosages accurately. Thus, the intricacies of alfalfa's genetic structure call for more sophisticated SNP genotyping systems capable of addressing the unique challenges posed by its autotetraploid nature. The investment cost and reliance upon skilled bioinformatics support for each genotyping run, make this a high-risk technology for breeders to adopt.

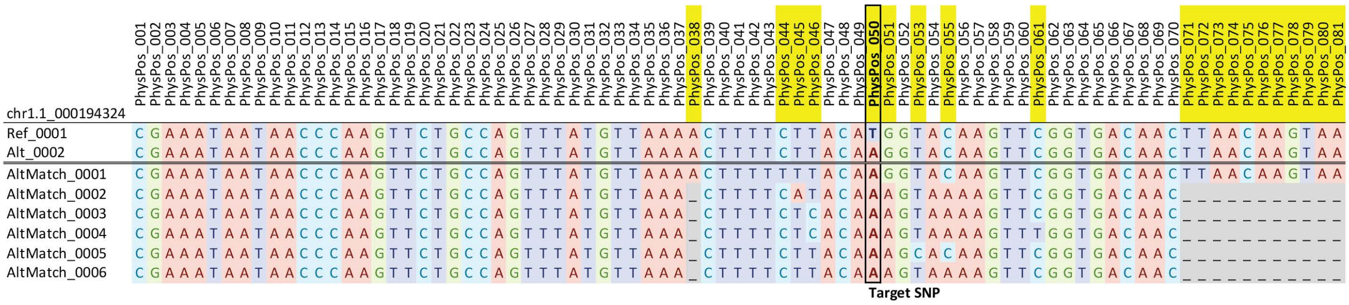
Currently, most alfalfa cultivars are synthetic populations developed by multiple cycles of phenotypic selection for desired traits. Evaluation for biomass yield, winter survival, grow back, disease resistance and forage nutritional quality among other traits is a multiyear process before cultivar registration and commercial seed production. Breeding programmes have operated with half-sib populations originating from polycrosses, where only the maternal parent is known, and the paternal parent can range from a few individuals up to hundreds. Unfortunately, breeding for yield gains in alfalfa using traditional phenotypic evaluation and recurrent selection methods has hit a plateau partly due to its highly heterozygous and heterogeneous population-level breeding. Community genomic tools like those in the 'Tools for Polyploids' project (https://www.polyploids.org/) or developed in other polyploid outcrossing crops have shown promise in accelerating breeding and yield gains (Ferrão *et al*, 2021). The creation of genomic tools that account for the biological and logistic challenges of the crop, has the potential to significantly improve yield gains in alfalfa through breeding.

The first and typically most tractable place to build capacity and tools for molecular breeding is to begin by creating a rapid genotyping pipeline that fits within both the breeding cycle and the selection cycle and can deliver on the breeder's objectives (Hawkins and Yu, 2018; Mejia-Guerra *et al*, 2021). As stated here, a pipeline refers to a complete workflow starting with a genetic marker platform, vendors for service and bioinformatic tools to transform returned raw data into a usable format for breeders. There are several factors to consider when choosing a genetic marker platform: cost per data point, vendor services, turnaround times and what genetic analyses can be done with the resulting data. For alfalfa, we determined that a targeted-amplicon sequenced-based approach would be the most beneficial for breeders. Unlike Genotyping-by-Sequencing (GBS), targeted, amplicon-based genotyping technologies such as DArTag (Diversity Array Technology - DArT), and Capture-Seq (LGC Genomics) have low missing data rates and query the same exact loci in all samples across genotyping projects, allowing new data to be easily appended to existing data (Telfer *et al*, 2019; Darrier *et al*, 2019; Wang *et al*, 2020). The amount of data returned is in the tens of thousands or less, rather than the millions of reads from GBS, simplifying downstream bioinformatics processing (Darrier *et al*, 2019; Milner *et al*, 2019). This in turn speeds up the analysis time for marker-assisted selection (MAS), introgression tracking, linkage mapping and genomic prediction (Darrier *et al*, 2019).

Here, we report the creation of a DArTag panel of 3,000 loci distributed across the alfalfa genome for use in molecular breeding and genomic prediction. DArTag is a hybridization/amplicon-based targeted genotyping platform developed by DArT (Blyton *et al* (2023); https://www.diversityarrays.com/services/targeted-genotoying/). Oligos are custom designed to target known genetic variants (SNPs and InDels less than 50bp) with its flanking genomic regions, and sequencing products of 54bp (legacy technology) or 81bp (current technology) in length are produced.

The DArTag assay consists of four steps based on principles described in Krishnakumar *et al* (2008). Briefly, the pool of 3,000 alfalfa oligos, each targeting one genetic variant, is hybridized to denatured gDNA in step 1, followed by SNP/INDEL copying into DArTag molecules by DNA polymerase in step 2. After ligation into circular molecule also in step 2, and nucleases treatment to remove unwanted molecules in step 3, DArTag products are subsequently amplified in step 4 with simultaneous addition of sample unique barcode used downstream for demultiplexing. The products of DArTag assay, after purification and quantification, are sequenced on NGS platforms (e.g. NovaSeq 6000, Illumina) to a depth of around 350x per marker per sample, then demultiplexed and genetic variants detected using DArT's proprietary analytical pipeline.

The alfalfa DArTag panel was designed on the legacy technology to produce 54bp reads but works equally well with the current technology (81bp reads) with the caveat that some residual adapter sequences may be included (read-through of the entire fragment into the adapter). After trimming of any residual adapter sequences, the reads can be used to call SNPs, or in the case of complex genomes like alfalfa, used to identify microhaplotypes (Figure 1). Sequencing reads can contain variants beyond the target SNP, which allows for the detection of more than two alleles at each of the 3,000 loci. As the amplicons are very short, variants found within these reads are assumed to be in complete linkage disequilibrium and therefore can be used for phasing genotyping calls for genetic map construction.



```
chr1.1_000194324
Ref_0001       C G A A A T A A T A A C C C A A G T T C T G C C A G T T T A T G T T A A A A C T T T T C T T A C A T G G T A C A A G T T C G G T G A C A A C T T A A C A A G T A A
Alt_0002       C G A A A T A A T A A C C C A A G T T C T G C C A G T T T A T G T T A A A A C T T T T C T T A C A A G G T A C A A G T T C G G T G A C A A C T T A A C A A G T A A
AltMatch_0001  C G A A A T A A T A A C C C A A G T T C T G C C A G T T T A T G T T A A A A C T T T T T T T A C A A G G T A C A A G T T C G G T G A C A A C T T A A C A A G T A A
AltMatch_0002  C G A A A T A A T A A C C C A A G T T C T G C C A G T T T A T G T T A A A _ C T T T T C A T A C A A A G T A A A A G T T C G G T G A C A A C _ _ _ _ _ _ _ _ _ _ _
AltMatch_0003  C G A A A T A A T A A C C C A A G T T C T G C C A G T T T A T G T T A A A _ C T T T T C T C A C A A A G T A A A A G T T C G G T G A C A A C _ _ _ _ _ _ _ _ _ _ _
AltMatch_0004  C G A A A T A A T A A C C C A A G T T C T G C C A G T T T A T G T T A A A _ C T T T T C T C A C A A A G T A A A A G T T T G G T G A C A A C _ _ _ _ _ _ _ _ _ _ _
AltMatch_0005  C G A A A T A A T A A C C C A A G T T C T G C C A G T T T A T G T T A A A _ C T T T T C T T A C A A A G C A C A A G T T C G G T G A C A A C _ _ _ _ _ _ _ _ _ _ _
AltMatch_0006  C G A A A T A A T A A C C C A A G T T C T G C C A G T T T A T G T T A A A _ C T T T T C T T A C A A A G T A A A A G T T C G G T G A C A A C _ _ _ _ _ _ _ _ _ _ _
```

Figure 1. DArTag sequencing reads from locus Chr1.1_000194324. Each sequence is a microhaplotype detected in breeding material tested on the panel. The DArTag assay was designed to detect the target locus (black rectangle) and distinguish the Reference allele from the Alternative allele. Additional variant nucleotide positions (yellow fill) distinguish the individual microhaplotypes. InDels are shown in grey fill. PhysPos refers to the physical nucleotide position within the sequencing read from left to right.

In addition, accurate allele dosage can be determined for both bi-allelic and multi-allelic haplotypes, allowing genetic effect contributions to be determined for each unique haplotype for traits of interest. As DArT had not tested many polyploid species with DArTag when this study was initiated, we agreed to limit the number of probes to 3,000 loci, though the optimal max may differ by species and genome complexity, and read depth required to sufficiently call genotypes (Andrzej Kilian, DArT, personal communication).

Results

This alfalfa 3K DArTag panel was developed from a diversity panel of 40 individual alfalfa clonal genotypes, focusing on elite breeding and stress-resistant genotypes used in North America. This panel consisted of 17 elite parents with various fall dormancy levels, six samples of diploid-cultivated alfalfa, 13 genotypes with abiotic stress resistance, one genotype with Aphanomyces root rot disease resistance, and three other genotypes (Table 1, column 2). Two biological replicates of the diversity panel were processed, where the sequencing libraries were prepared using either Illumina Nextera WGS library prep at Cornell Institute of Biotechnology or NEBNext Ultra DNA Library Prep Kit with an average insert DNA size of 300bp. The whole-genome sequencing (WGS) was done using Illumina NovaSeq 6000 at Novogene (https://en.novogene.com/). Raw FASTQ sequences were processed by remov-ing residual adapter sequences and low-quality bases using Trimmomatic (LEADING:10 TRAILING:10 SLIDINGWINDOW:4:15 MINLEN:30) (Bolger *et al*, 2014). Cleaned reads were then aligned to the haploid set (the first set out of the four homologous chromosomes) of XinJiangDaYe reference genome (Chen *et al*, 2020) using BWA-MEM (Li, 2012, 2013) and structural variants (SNPs and indels) were called using the DNAseq pipeline developed by Sention (https://www.sentieon.com). A total of 28M SNPs present in both replicates were discovered from the whole-genome re-sequencing of the diversity panel, where a high-confidence set of 10K SNPs (Figure 2) were obtained by requiring them: (1) not located within 5bp distance to an indel, (2) QUAL $>$ 30, (3) minimum and maximum read depths of 20 and 1,900, respectively, (4) for each sample, at least one read supporting reference allele and two reads supporting the alternative allele, (5) no missing genotype per SNP position, (6) with a minor allele frequency greater than 0.25, (7) not located in transposable elements and (8) not within 1Kb of chromosome termini. The 10K SNPs were assessed by DArT, and from those that passed QC, a 3K SNP set targeting even genomic distribution was selected to form a 3K DArTag marker panel. Of the 3,000 loci selected for the panel, 85% (2,542) reside in genic regions and only 15% (458) reside in non-genic regions (Supplemental Table 1). Oligo probes were synthesized, and genotyping done at DArT.

The alfalfa 3K DarTag marker panel was validated using a bi-parental F1 population (n = 184), a backcross (BC1) population (n = 94), and a diverse set of elite genotypes (n = 74) and individual plants from other *Medicago* species (n = 20) (Table 1, column 3). It should be noted that all 40 alfalfa lines used in the SNP discovery were also included in this validation sample set. The material selected for validation was to assess (1) the panel's ability to construct genetic (linkage) maps with the data output and (2) to define the usable limit to the panel with extant species (non-*Medicago sativa*) germplasm.

As expected, the missing data (a marker with $<$ 10 reads in a population) is the lowest among the *Medicago sativa* genotypes. The alfalfa lines used in SNP discovery showed the least missing data (an average of ~9% of the 3K markers with no data) and the rest *M. sativa* lines of the validation sample set had comparable missing rates (10%) (Supplemental Figure 2). Other *Medicago* species had an average of 51% markers with missing data, which is approximately five times higher than the *M. sativa* genotypes.

Table 1. Accessions used in the construction and testing of the alfalfa (*Medicago sativa* L.) 3K DArTag panel. Germplasm used for whole genome sequencing and SNP database construction are indicated by 'Y' in the 'SNP discovery' column. Germplasm used to validate the 3K panel is indicated by 'Y' in the 'Validation set' column.

Sample ID	SNP discovery	Validation set	Contributor	Note
S&W FD4	Y	Y	S&W Seed Co.	Elite parent; fall dormancy 4
S&W FD5	Y	Y	S&W Seed Co.	Elite parent; fall dormancy 5
Legacy FD4	Y	Y	Legacy Seeds	Elite parent; fall dormancy 4
Legacy FD5	Y	Y	Legacy Seeds	Elite parent; fall dormancy 5
S&W FD6	Y	Y	S&W Seed Co.	Elite parent; fall dormancy 6
S&W FD7	Y	Y	S&W Seed Co.	Elite parent; fall dormancy 7
S&W FD8	Y	Y	S&W Seed Co.	Elite parent; fall dormancy 8
S&W FD9A	Y	Y	S&W Seed Co.	Elite parent; fall dormancy 9
S&W SFD9B	Y	Y	S&W Seed Co.	Elite parent; fall dormancy 9
CADL-1	Y	Y	N. Young	Cultivated alfalfa at diploid level
CADL-3	Y	Y	N. Young	Cultivated alfalfa at diploid level
CADL-4-5	Y	Y	N. Young	Cultivated alfalfa at diploid level
CADL-5-3	Y	Y	N. Young	Cultivated alfalfa at diploid level
CADL-13	Y	Y	N. Young	Cultivated alfalfa at diploid level
CADL-18	Y	Y	N. Young	Cultivated alfalfa at diploid level
UMN3988-BIP	Y	Y	D. Samac	Biomass type
RegenSY27x	Y	Y	D. Samac	Regenerator; Ref. genome
I195	Y	Y	N. Young	WAPH5; Aphanomyces root rot
UT14-46 SP	Y	Y	M. Peel	Tetraploid *Medicago falcata*
UT27-62	Y	Y	M. Peel	Elite parent; Salt tolerant
FL99	Y	Y	E. Rios	Elite parent; Fall dormancy 9
Bulldog 505	Y	Y	A. Missaoui	Elite parent; fall dormancy 5
GAMS 1403-FSH	Y	Y	A. Missaoui	Elite parent; fall dormancy 7
GAMS 1404-FSH	Y	Y	A. Missaoui	Elite parent; fall dormancy 8
GAMS 1405-FSH	Y	Y	A. Missaoui	Elite parent; fall dormancy 9
3010	Y	Y	A. Missaoui	Elite parent; fall dormancy 3
CW1010	Y	Y	A. Missaoui	Elite parent; fall dormancy 10
CUF 101	Y	Y	D. Samac	Fall dormancy 10 check
BIP1	Y	Y	M. Peel	Salt tolerant; 27-62
BIP2	Y	Y	M. Peel	Salt tolerant; 31-6
BIP3	Y	Y	M. Peel	SemiP; 1-34
BIP4	Y	Y	M. Peel	SemiP; 6-2
BIP5	Y	Y	M. Peel	SemiP; 14-46
BIP6	Y	Y	M. Peel	Drought (Ut7); 17-43
BIP7	Y	Y	M. Peel	Drought (Ut8); 17-44
BIP8	Y	Y	M. Peel	Drought (Ut9); 18-22
BIP9	Y	Y	M. Peel	Drought (Ut10); 21-3
BIP10	Y	Y	M. Peel	Drought (Ut11); 22-30
BIP11	Y	Y	M. Peel	Drought (Ut26); 7-18
BIP12	Y	Y	M. Peel	Drought (Ut30); 13-14
Wilson	N	Y	L.-X. Yu	Elite parent
WA467895	N	Y	L.-X. Yu	Elite parent
Cornell NY1	N	Y	D. Viands	Elite parent
Cornell NY2	N	Y	D. Viands	Elite parent
Cornell NY3	N	Y	D. Viands	Elite parent
Cornell NY4	N	Y	D. Viands	Elite parent
PAF 13 5, 11-1	N	Y	H. Riday	*Medicago falcata*

Continued on next page

Table 1 continued

Sample ID	**SNP discovery**	**Validation set**	**Contributor**	**Note**
PAF 13 2, 9-4	N	Y	H. Riday	*Medicago falcata*
PAF 13 9, 10-5	N	Y	H. Riday	*Medicago falcata*
PAF 13 7, 21-2	N	Y	H. Riday	*Medicago falcata*
FAL12 1, 11-5	N	Y	H. Riday	*Medicago falcata*
FAL12 4, 12-4	N	Y	H. Riday	*Medicago falcata*
MAV8	N	Y	D. Samac	Elite parent
Aph 2	N	Y	D. Samac	Elite parent
MAV13	N	Y	D. Samac	Elite parent
MAV14	N	Y	D. Samac	Elite parent
MAV15	N	Y	D. Samac	Elite parent
ZG9	N	Y	D. Samac	Elite parent
ZG20	N	Y	D. Samac	Elite parent
ZG21	N	Y	D. Samac	Elite parent
ZG23	N	Y	D. Samac	Elite parent
ZG25	N	Y	D. Samac	Elite parent
Aph 11	N	Y	D. Samac	Elite parent
Aph 47	N	Y	D. Samac	Elite parent
PI 516640	N	Y	B. Irish	*Medicago arabica*
PI 504540	N	Y	B. Irish	*Medicago arborea*
PI 495215	N	Y	B. Irish	*Medicago bonarotiana*
PI 315458	N	Y	B. Irish	*Medicago cancellata*
PI 498767	N	Y	B. Irish	*Medicago ciliaris*
W6 32886	N	Y	B. Irish	*Medicago daghestanica*
PI 538998	N	Y	B. Irish	*Medicago hybrida*
PI 498849	N	Y	B. Irish	*Medicago laciniata*
PI 537186	N	Y	B. Irish	*Medicago littoralis*
PI 516711	N	Y	B. Irish	*Medicago marina*
PI 287999	N	Y	B. Irish	*Medicago monspelliaca*
PI 537259	N	Y	B. Irish	*Medicago murex*
PI 220021	N	Y	B. Irish	*Medicago orbicularis*
PI 464704	N	Y	B. Irish	*Medicago papillosa*
PI 253450	N	Y	B. Irish	*Medicago pironae*
W6 5252	N	Y	B. Irish	*Medicago polymorpha*
PI 150564	N	Y	B. Irish	*Medicago popovii*
PI 577446	N	Y	B. Irish	*Medicago prostrata*
PI 631912	N	Y	B. Irish	*Medicago ruthenica*
PI 631715	N	Y	B. Irish	*Medicago sativa nothosubsp. tunetana*
PI 631714	N	Y	B. Irish	*Medicago sativa nothosubsp. tunetana*
PI 631952	N	Y	B. Irish	*Medicago sativa nothosubsp. varia*
PI 631920	N	Y	B. Irish	*Medicago sativa nothosubsp. varia*
PI 631923	N	Y	B. Irish	*Medicago sativa subsp. caerulea*
PI 631921	N	Y	B. Irish	*Medicago sativa subsp. caerulea*
PI 641405	N	Y	B. Irish	*Medicago sativa subsp. glomerata*
PI 631978	N	Y	B. Irish	*Medicago sativa subsp. glomerata*
PI 631869	N	Y	B. Irish	*Medicago sativa var. viscosa*
PI 631870	N	Y	B. Irish	*Medicago sativa var. viscosa*
PI 197356	N	Y	B. Irish	*Medicago scutellata*
I195 x J432	N	Y	D. Samac	F1 population (184 progeny)
AphBC1	N	Y	D. Samac	BC1 population (94 progeny)

Using the 3K panel genotyping results, we generated linkage maps for two distinct populations, an F1 and a backcross (BC1) that share the parent I195. For the F1 population, individuals were derived from a cross between parents I195 and J432, which are resistant and susceptible to *Aphamomyces euteiches*, respectively. Meanwhile, the BC1 population was obtained through a cross between I195 and a progeny (85-209) from the above F1 population. Initially, we constructed individual genetic maps for the F1 and BC1 populations. Genotype dosages for both were determined using updog software (Gerard *et al*, 2018). Subsequently, updog-generated objects were fed into MAPpoly software (Mollinari and Garcia, 2019; Mollinari *et al*, 2020) to build separate genetic maps for each population. A standard screening was performed based on missing data and Mendelian segregation fit. We calculated the recombination fraction matrix between all retained markers, using this information to cluster markers into linkage groups. According to available genome information, most of these markers corresponded with specific chromosomes. Notably, a few markers that were mapped outside their physical position still presented consistent linkage with the markers in their assigned group. This pattern held true across both F1 and BC1 populations. For each linkage group formed, we used MAPpoly's functions, *mds_mappoly* and *est_rf_hmm_sequential* to carry out de novo ordering and phasing to obtain the final F1 and BC1 maps. From all mapped markers, only 2.55% were assigned to different chromosomes in the F1 map and 0.97% in the BC1 map. After constructing individual maps for the F1 and BC1 populations, we merged them using the genome order (Figure 3A; Supplemental Figure 1A). The mapped markers were all consistent placed in the two maps, but a few markers were assigned to different linkage groups when comparing the linkage and physical assembly in both maps. These markers were retained in filtering because they had reasonable Mendelian segregation behaviour and their association with linkage groups that do not correspond to their physical chromosome assignment could indicate potential errors in the reference assembly (Figure 3B). Markers mapped out of their physical positions were inserted into the genome-based map using the multidimensional scaling (MDS) de novo information. We then reconstructed a joint map by employing the hidden Markov model (HMM) algorithm's extension, as Mollinari and Garcia (2019) detailed. The implementation for this algorithm can be found in the GitHub repository https://github.com/mmollina/highprecHMM. Finally, haplotypes for all individuals across both F1 and BC1 populations were reconstructed using the same algorithm (Supplemental Figure 1B).

Conclusion

This panel is now publicly available and open for any researcher or breeder to order through DArT (https://www.diversityarrays.com). Researchers interested in using the panel and genotyping services are encouraged to contact DArT directly for pricing details.

Raw data in FASTQ can be requested as can the Missing Allele Discovery File (MADC) that indicates the read depth of each detected haplotype in each sample. The panel and its resulting data are suitable for marker-assisted selection, reconstruction of recombination patterns, allele dosage estimation, and parental confirmation in North American cultivated alfalfa, with some limited application in other *Medicago* species. The efficacy of the panel on breeding materials outside of North America has not been tested, nor has its efficacy in GWAS. Single plant samples were used to create and test the panel. Subsequent testing on samples that are genotyped individually and in tissue or DNA bulks (DNA bulks up to 30 individuals per population) have produced the same allele frequency ratios in both sample types but higher read depth in pools (Esteban Rios, personal communication). More testing is needed to determine the most efficient number of samples to pool to achieve population-level allele frequencies with minimal human labour and monetary costs.

The DArTag assay can be processed from gDNA or from tissue to genotyping data extraction in a three-week turnaround time. The genotyping data report comprises allele dose calls and raw data with custom report formats available upon request. One benefit that DArTag has over fixed array platforms is the ability to update and improve the panel as required over time. The panel is a pool of 3,000 oligos, one per locus, which is used to generate the sequencing libraries from the assayed material. Because the pool is created from individual oligo stocks, the removal of suboptimal loci or the addition of new loci can be easily done by creating a new pool. To determine which loci should be considered for removal, extensive genotyping (> 10,000 samples) is underway to identify those loci that consistently underperform or fail and flag them for removal. Independently, as new significant QTL markers and/or markers specific to other germplasm are detected, they can be targeted for inclusion in the original pool in the next version(s) of the panel. DArT offers re-pooling services once per year at low or no cost, but more frequent requests could result in labour surcharges being applied (Andrzej Kilian, personal communication). Researchers interested in initiating projects with DArT are encouraged to contact DArT directly for consultation.

Another benefit of the deep testing underway is the ability to detect and catalogue all the microhaplotypes into a fixed allele database, which will improve combining data sets across genotyping projects (manuscript in preparation). If after deep testing it is clear that there are too few markers for GWAS for given traits of interest, additional panels can be made to complement this panel. The other option is to add the required loci to the existing panel up to the technical limit of 7K, which is

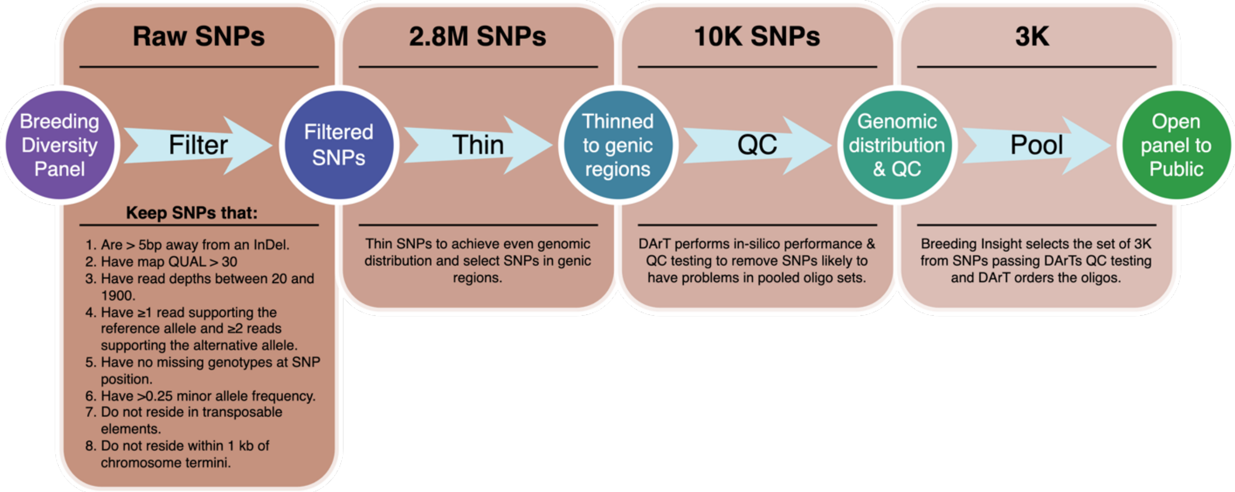

Raw SNPs
2.8M SNPs
10K SNPs
3K
Breeding Diversity Panel
Filter
Filtered SNPs
Thin
Thinned to genic regions
QC
Genomic distribution & QC
Pool
Open panel to Public
Keep SNPs that:
1. Are > 5bp away from an InDel.
2. Have map QUAL > 30
3. Have read depths between 20 and 1900.
4. Have ≥1 read supporting the reference allele and ≥2 reads supporting the alternative allele.
5. Have no missing genotypes at SNP position.
6. Have >0.25 minor allele frequency.
7. Do not reside in transposable elements.
8. Do not reside within 1 kb of chromosome termini.
Thin SNPs to achieve even genomic distribution and select SNPs in genic regions.
DArT performs in-silico performance & QC testing to remove SNPs likely to have problems in pooled oligo sets.
Breeding Insight selects the set of 3K from SNPs passing DArTs QC testing and DArT orders the oligos.

Figure 2. Filters and criteria applied to produce the 3K DArTag SNP panel from the whole-genome sequencing (WGS) of the alfalfa diversity panel. M, millions; K, thousands.

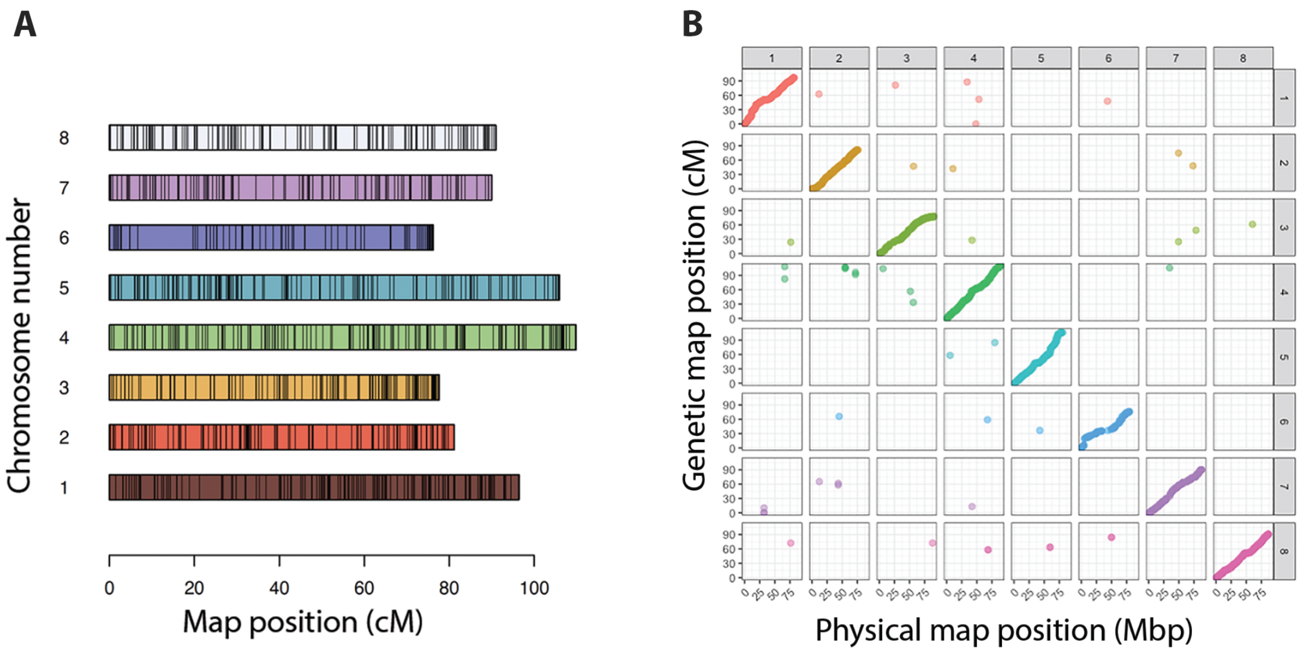


Figure 3. Composite genetic maps of a bi-parental F1 and a backcross (BC1) population. A) Regeneration of the eight linkage groups of alfalfa genome. Scale bar is shown in cM. B) Scatter plots showing the relationship of genetic distance (cM) to physical distance (Mbp) for each of the eight linkage groups.

a more cost-effective option for the routine genotyping service with scalability.

We choose to create a panel of 3,000 loci due to cost and technical reasons, but smaller complementary panels can be made at lower up-front and downstream usage costs. The addition of a complementary 3K panel would nearly double the cost of genotyping per sample but would result in more granular genotyping data.

Data availability statement

The FASTQ files from the whole-genome skim sequencing for the 40 *Medicago sativa* accessions used for identifying the candidate SNP variants are housed in the NCBI Short Read Archive under the BioProject ID PRJNA1014379. The targeted regions used to create the 3K DArTag markers and the haplotypes detected as of 31 May 2023 (v17) are available on DRYAD (https://datadryad.org/stash/share/wJEn32Dfl94EOYMoeM00PJti6MKUliPBTAtsgbWJyOU). The code and data for construction of the F1, BC1 and joint maps in MAPpoly are available in our GitHub repository for those interested in reproducing our analysis (https://github.com/Breeding-Insight/alfalfa_dartag_panel_paper.git).

Acknowledgments

Breeding Insight is acknowledged for project design, marker development, curation and data processing. Diversity Arrays Technology created the oligo array, provided sequencing services and contributed to the manuscript. We thank S&W Seed Company, Legacy Seeds, Debby Samac, Brian Irish, Michael Peel, Long-Xi Yu, Heathcliffe Riday, Nevin Young, Esteban Rios, Ian Ray, Ali Missaoui and Don Viands for providing germplasm and Sharon Mitchell for manuscript review. Breeding Insight was funded for this work through a Cooperative Agreement between USDA-ARS and Cornell (project number: 8062-21000-043-004-A). Marcelo Mollinari was funded by a USDA NIFA-awarded AFRI grant (project number: 2022-67013-36269).

Supplemental Data

Supplemental Figure 1. Alfalfa Genetic map construction for an F1, BC1, and a joint map of the consensus.
Supplemental Figure 2. Missing data rates for different grouped subsets of genetic material.
Supplemental Table 1. Final 3,000 loci selected for the DArTag panel.

Author contributions

DZ, KMMG, DS and MJS contributed to experimental design and planning. DZ, DS and MJS selected the diversity panel for WGS. DS, MP and BI grew and harvested all plant materials used in the study. KMMG performed all the WGS analyses, SNP database creation, filtering pipelines and quality control analyses to create the 3K panel. KHU managed the panel creation at Diversity Arrays Technology. DZ, MM and DS executed the data analyses and genetic mapping. DZ and MJS wrote the initial draft of the manuscript. CB managed experiments and communication among all authors involved. All authors contributed to reviewing the manuscript.

Conflict of interest statement

The authors have no conflicts of interest to report.

References

Blyton, M. D. J., Brice, K. L., Heller-Uszynska, K., Pascoe, J., Jaccoud, D., Leigh, K. A., and Moore, B. D. (2023). A new genetic method for diet determination from faeces that provides species level resolution in the koala. *bioRxiv* 2023.02.12.528172. doi: https://doi.org/10.1101/2023.02.12.528172

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinform* 30(15), 2144–2120. doi: https://doi.org/10.1093/bioinformatics/btu170

Chen, H., Zeng, Y., Yang, Y., Huang, L., Tang, B., Zhang, H., Hao, F., Liu, W., Li, Y., Liu, Y., Zhang, X., Zhang, R., Zhang, Y., Li, Y., Wang, K., He, H., Wang, Z., Fan, G., Yang, H., Bao, A., Shang, Z., Chen, J., Wang, W., and Qiu, Q. (2020). Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa. *Nature Comm* 11, 1–11. doi: https://doi.org/10.1038/s41467-020-16338-x

Darrier, B., Russell, J., Milner, S. G., Hedley, P. E., Shaw, P. D., Macaulay, M., Ramsay, L. D., Halpin, C., Mascher, M., Fleury, D. L., Landridge, P., Stein, N., and Waugh, R. (2019). A Comparison of Mainstream Genotyping Platforms for the Evaluation and Use of Barley Genetic Resources. *Front Plant Sci* 10, 1–14. doi: https://doi.org/10.3389/fpls.2019.00544

Eathington, S. R., Crosbie, T. M., Edwards, M. D., Reiter, R. S., and Bull, J. K. (2007). Molecular markers in a commercial breeding program. *Crop Sci* 47, 154–163. doi: https://doi.org/10.2135/cropsci2007.04.0015IPBS

Fernandez, A. L., Sheaffer, C. C., Tautges, N. E., Putnam, D. H., and Hunter, M. C. (2019). Alfalfa, Wildlife, and the Environment (St. Paul, MN: National Alfalfa and Forage Alliance).

Ferrão, L. F. V., Amadeu, R. R., Benevenuto, J., Oliveira, I. D. B., and Munoz, P. R. (2021). Genomic Selection in an Outcrossing Autotetraploid Fruit Crop: Lessons From Blueberry Breeding. *Front Plant Sci* 12, 1–13. doi: https://doi.org/10.3389/fpls.2021.676326

Feurerstein, U., Brown, A. H. D., and Burdon, J. J. (1990). Linkage of Rust Resistance Genes from Wild Barley (*Hordeum spotaneum*) with Isozyme Markers. *Plant Breeding* 104, 318–324. doi: https://doi.org/10.1111/j.1439-0523.1990.tb00442.x

Gerard, D., Ferrão, L. F. V., Garcia, A. A. F., and Stephens, M. (2018). Genotyping Polyploids from Messy Sequencing Data. *Genetics* 210(3), 789–807. doi: https://doi.org/10.1534/genetics.118.301468

Hasan, N., Choudhary, S., Naaz, N., Sharma, N., and Laskar, R. A. (2021). Recent advancements in molecular marker-assisted selection and applications in plant breeding programmes. *J Genet Eng Biotech* 19, 1–26. doi: https://doi.org/10.1186/s43141-021-00231-1

Hawkins, C. and Yu, L. X. (2018). Recent progress in alfalfa (*Medicago sativa* L.) genomics and genomic selection. *The Crop Journal* 6, 565–575. doi: https://doi.org/10.1016/j.cj.2018.01.006

Heffner, E. L., Sorrells, M. E., and Jannink, J. L. (2009). Genomic selection for crop improvement. *Crop Sci* 49, 1–12. doi: https://doi.org/10.2135/cropsci2008.08.0512

Helentjaris, T., King, G., Slocum, M., Siedenstrang, C., and Wegman, S. (1985). Restriction fragment polymorphisms as probes for plant diversity and their development as tools for applied plant breeding. *Plant*

Mol Biol 5, 109–118. doi: https://doi.org/10.1007/BF00020093

Krishnakumar, S., Zheng, J., Wilhelmy, J., Faham, M., Mindrinos, M., and Davis, R. (2008). A comprehensive assay for targeted multiplex amplification of human DNA sequences. *PNAS* 105(27), 9296–9301. doi: https://doi.org/10.1073/pnas.0803240105

Li, H. (2012). Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. *Bioinform* 28(14), 1838–1882. doi: https://doi.org/10.1093/bioinformatics/bts280

Li, H. (2013). Aligning sequence reads, clones sequences and assembly contigs with BWA-MEM. *arXiv* 1303.3997. doi: https://doi.org/10.48550/arXiv.1303.3997

Lorenzana, R. and Bernardo, R. (2009). Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. *Theor Appl Genet* 120, 151–161. doi: https://doi.org/10.1007/s00122-009-1166-3

Mejia-Guerra, M. K., Zhao, D., and Sheehan, M. J. (2021). Genomic Resources for Breeding in Alfalfa: Availability, Utility, and Adoption. In *The Alfalfa Genome, Compendium of Plant Genomes,* ed. Yu, L. X. and Kole, C., (Cham: Springer), 177-189.

Milner, S. G., Jost, M., Taketa, S., Mazon, E. R., Himmelbach, A., Oppermann, M., Weise, S., Knupffer, H., Basterrechea, M., König, P., Schüler, D., Sharma, R., Pasam, R. K., Rutten, T., Guo, G., Xu, D., Zhang, Z., Herren, G., Müller, T., Krattinger, S. G., Keller, B., Jiang, Y., González, M. Y., Zhao, Y., Habekuß, A., Fäber, S., Ordon, F., Lange, M., Börner, A., Graner, A., Reif, J. C., Scholz, U., Mascher, M., and Stein, N. (2019). Genebank genomics reveals the diversity of a global barley collection. *Nat. Genet* 51, 319–326. doi: https://doi.org/10.1038/s41588-018-0266-x

Mollinari, M. and Garcia, A. A. F. (2019). Linkage Analysis and Haplotype Phasing in Experimental Autopolyploid Populations with High Ploidy Level Using Hidden Markov Models. *Genes, Genomes, Genetics* 3(10), 3297–3314. doi: https://doi.org/10.1534/g3.119.400378

Mollinari, M., Olukolu, B. A., Gds, P., Khan, A., Gemenet, D., Yencho, G. C., and Zeng, Z. (2020). Unraveling the Hexaploid Sweetpotato Inheritance Using Ultra-Dense Multilocus Mapping. *Genes, Genomes, Genetics* 3(1), 281–292. doi: https://doi.org/10.1534/g3.119.400620

Putnam, D. and Meccage, E. (2022). Profitable alfalfa production sustains the environment. In Proceedings, 2022 World Alfalfa Congress, 14-17 November 2022, San Diego, CA.

Tanksley, S. D. (1983). Molecular markers in plant breeding. *Plant Mol Biol Rep* 1, 3–8. doi: https://doi.org/10.1101/2023.02.12.528172

Telfer, E., Graham, N., Macdonald, L., Li, Y., Klápště, J., Resende, M., Neves, L. G., Dungey, H., and Wilcox, P. (2019). A high-density exome capture genotype-by-sequencing panel for forestry breeding in Pinus radiata. *PLoS One* 14(9), 222640–222640. doi: https://doi.org/10.1371/journal.pone.0222640

Undersander, D. (2021). Economic importance, practical limitations to production, management, and breeding targets of alfalfa. In *The Alfalfa Genome, Compendium of Plant Genomes,* ed. Yu, L. X. and Kole, C., (Cham: Springer), 1-11.

Wang, N., Yuan, Y., Wang, H., Yu, D., Liu, Y., Zhang, A., Gowda, M., Nair, S. K., Hao, Z., Lu, Y., Vincente, F. S., Prasanna, B. M., Li, X., and Zhang, X. (2020). Applications of genotyping-by-sequencing (GBS) in maize genetics and breeding. *Sci Rep* 10. doi: https://doi.org/10.1038/s41598-020-73321-8